

New Stability Results and Algorithms for Block Tridiagonal Systems, with Applications to Kalman Smoothing

Aleksandr Y. Aravkin, Bradley M. Bell, James V. Burke, Gianluigi Pillonetto

Abstract—Block tridiagonal systems play a key role in all Kalman smoothing applications, including the classic Rauch-Tung-Striebel smoother, as well as more modern variants that incorporate nonlinear models, inequality constraints on the state, and robust penalties on state and measurement components.

The paper begins with a fundamental result that connects the condition number of a common block tridiagonal system to properties of the individual blocks. As a consequence, we obtain sufficient conditions for the stability of Kalman smoothing formulations. The result also illustrates how unstable Kalman smoothing problems can arise, and *how to easily stabilize them*.

Then, turning our attention to algorithms, we show that the classic Rauch-Tung-Striebel smoother is an implementation of the forward Thomas algorithm for block tridiagonal systems. We reveal a flaw in the existing theory for characterizing the numerical stability of the algorithm, and provide a new numerical stability result that shows the condition number of every recursively modified block is bounded by the condition number of the whole system as the algorithm proceeds.

Finally, we present a new backward block tridiagonal Thomas algorithm, and show that the condition numbers of the recursively modified blocks in this algorithm are independent of the condition number of the full block tridiagonal system.

I. INTRODUCTION

Kalman filtering and smoothing methods form a broad category of computational algorithms used for inference on noisy dynamical systems. Since their invention [15] and early development [16], these algorithms have become a gold standard in a range of applications, including space exploration, missile guidance systems, general tracking and navigation, and weather prediction. In 2009, Rudolf Kalman received the National Medal of Science from President Obama for the invention of the Kalman filter. Numerous books and papers have been written on these methods and their extensions, addressing modifications for use in nonlinear systems [14], robustification against unknown models [22], smoothing data over time intervals [1], [6], Kalman smoothing with unknown parameters [8], constrained Kalman smoothing [7], and improving robustness to bad measurements [11], [21], [10], [24], [20], [18], [17], [5] and sudden changes in state [3], [4], [13].

Aleksandr Y. Aravkin (saravkin@us.ibm.com) is with the IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

Bradley M. Bell (bradbell@apl.washington.edu) is with Applied Physics Lab University of Washington, Seattle, WA

James V. Burke (burke@math.washington.edu) is with Department of Mathematics University of Washington, Seattle, WA

Gianluigi Pillonetto (giapi@dei.unipd.it) is with Dipartimento di Ingegneria dell'Informazione, University of Padova, Padova, Italy.

Therefore, we use the term ‘Kalman smoothing’ to refer to any system that can be represented graphically by Figure 1. In order to motivate the analysis in this paper, we begin by reviewing the classic Kalman smoothing problem, and show that it reduces to solving a linear system with block tridiagonal structure.

The model corresponding to Figure 1 is specified as follows:

$$\begin{aligned} x_1 &= g_1(x_0) + w_1, \\ x_k &= g_k(x_{k-1}) + w_k \quad k = 2, \dots, N, \\ z_k &= h_k(x_k) + v_k \quad k = 1, \dots, N, \end{aligned} \quad (1.1)$$

where w_k, v_k are mutually independent random variables with known positive definite covariance matrices Q_k and R_k , respectively. Extensions have been proposed that work with singular Q_k and R_k (see e.g. [19], [9]), but in this paper we confine our attention to the classic nonsingular case.

We have $x_k, w_k \in \mathbb{R}^n$, and $z_k, v_k \in \mathbb{R}^{m(k)}$, so measurement dimensions can vary between time points. The classic case is obtained by making the following assumptions:

- 1) x_0 is known, and g_k, h_k are known linear functions, which we denote by

$$g_k(x_{k-1}) = G_k x_{k-1} \quad h_k(x_k) = H_k x_k \quad (1.2)$$

where $G_k \in \mathbb{R}^{n \times n}$ and $H_k \in \mathbb{R}^{m(k) \times n}$,

- 2) w_k, v_k are mutually independent Gaussian random variables.

If the models are linear and the noises are Gaussian, using Bayes’ theorem, we have

$$\mathbf{p}(x_k | z_k) \propto \mathbf{p}(z_k | x_k) \mathbf{p}(x_k) = \mathbf{p}(v_k) \mathbf{p}(w_k), \quad (1.3)$$

and therefore the likelihood of the entire state sequence $\{x_k\}$ given the entire measurement sequence $\{z_k\}$ is proportional to

$$\begin{aligned} \prod_{k=1}^N \mathbf{p}(v_k) \mathbf{p}(w_k) &\propto \prod_{k=1}^N \exp \left(-\frac{1}{2} (z_k - H_k x_k)^\top R_k^{-1} (z_k - H_k x_k) \right. \\ &\quad \left. - \frac{1}{2} (x_k - G_k x_{k-1})^\top Q_k^{-1} (x_k - G_k x_{k-1}) \right). \end{aligned} \quad (1.4)$$

A better (equivalent) formulation to (1.4) is minimizing its negative log posterior:

$$\begin{aligned} \min_{\{x_k\}} f(\{x_k\}) &:= \sum_{k=1}^N \frac{1}{2} (z_k - H_k x_k)^\top R_k^{-1} (z_k - H_k x_k) \\ &\quad + \frac{1}{2} (x_k - G_k x_{k-1})^\top Q_k^{-1} (x_k - G_k x_{k-1}). \end{aligned} \quad (1.5)$$

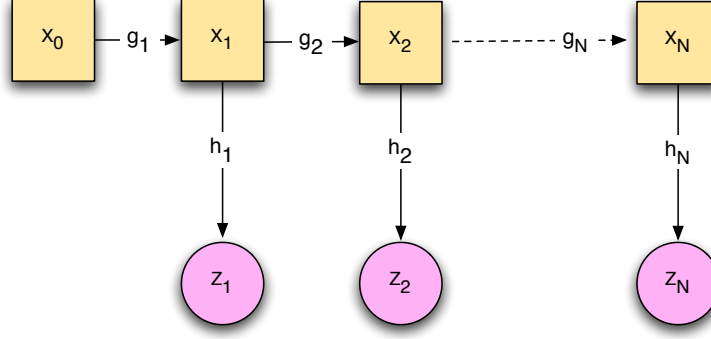


Fig. 1. Dynamic systems amenable to Kalman smoothing methods.

To simplify the problem, we now introduce data structures that capture the entire state sequence, measurement sequence, covariance matrices, and initial conditions.

Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

$$\text{vec}(\{v_k\}) = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}, \quad \text{diag}(\{T_k\}) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

We make the following definitions:

$$\begin{aligned} R &= \text{diag}(\{R_k\}) & x &= \text{vec}(\{x_k\}) \\ Q &= \text{diag}(\{Q_k\}) & \zeta &= \text{vec}(\{x_0, 0, \dots, 0\}) \\ H &= \text{diag}(\{H_k\}) & z &= \text{vec}(\{z_1, z_2, \dots, z_N\}) \end{aligned} \quad (\text{I.6})$$

$$G = \begin{bmatrix} I & 0 & & \\ -G_2 & I & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & I \end{bmatrix}. \quad (\text{I.7})$$

With definitions in (I.6) and (I.7), problem (I.5) can be written

$$\min_x f(x) = \frac{1}{2} \|Hx - z\|_{R^{-1}}^2 + \frac{1}{2} \|Gx - \zeta\|_{Q^{-1}}^2, \quad (\text{I.8})$$

where $\|a\|_M^2 = a^\top M a$. It is well known that the MAP was a least-squares problem, but this derivation makes the structure fully transparent. We now write down the linear system that needs to be solved in order to find the solution to this least-squares problem:

$$(H^\top R^{-1} H + G^\top Q^{-1} G)x = H^\top R^{-1} z + G^\top Q^{-1} \zeta. \quad (\text{I.9})$$

The linear system in (I.9) has a very special structure: it is symmetric positive definite block tridiagonal, since G is nonsingular and Q is positive definite. Direct computation shows that

$$C = (H^\top R^{-1} H + G^\top Q^{-1} G) = \begin{bmatrix} D_1 & A_2^\top & 0 & \\ A_2 & D_2 & A_3^\top & 0 \\ 0 & \ddots & \ddots & \ddots \\ 0 & 0 & A_N & D_N \end{bmatrix}, \quad (\text{I.10})$$

with $A_k \in \mathbb{R}^{n \times n}$ and $D_k \in \mathbb{R}^{n \times n}$ defined as follows:

$$\begin{aligned} A_k &= -Q_k^{-1} G_k, \\ D_k &= Q_k^{-1} + G_{k+1}^\top Q_{k+1}^{-1} G_{k+1} + H_k^\top R_k^{-1} H_k. \end{aligned} \quad (\text{I.11})$$

This block tridiagonal structure was noted early on in [25], [12], [26]. These systems also arise in many recent extended formulations, see e.g. [8, (15)], [7, (12)], [4, (8.13)].

The paper proceeds as follows. Section II focuses on the subclass of block tridiagonal systems that capture all of the systems of interest to various Kalman smoothing applications. Bounds on the eigenvalues of these systems are obtained in terms of the behavior of the individual components. It is shown how these bounds are related to the numerical stability of Kalman smoothing formulations. In Section III it is shown that the classic RTS smoother implements the forward block tridiagonal Thomas algorithm for the system (I.9). We demonstrate a flaw in the stability analysis for this algorithm given in [8], and prove a new more powerful stability result that shows *any* stable system from Section II can be solved using the forward block Thomas algorithm. Finally, in Section V, we introduce a new backward Thomas algorithm, and show that it behaves stably *independent of the condition number* of the full system. We conclude with a discussion of the consequences of the new results.

II. CHARACTERIZING BLOCK TRIDIAGONAL SYSTEMS

Consider systems of form

$$B = a^\top q a \quad (\text{II.1})$$

where

$$q = \text{diag}\{q_1, \dots, q_N\}, \quad a = \begin{bmatrix} I & 0 & & \\ a_2 & I & \ddots & \\ & \ddots & \ddots & 0 \\ & & a_N & I \end{bmatrix}.$$

Let λ_{\min} , λ_{\max} , and σ_{\min} , σ_{\max} denote the minimum and maximum eigenvalues and singular values, respectively. Simple upper bounds on the lower and upper eigenvalues of B are easily obtained.

Theorem 2.1:

$$\lambda_{\min}(q)\sigma_{\min}^2(a) \leq \lambda_{\min}(B) \leq \lambda_{\max}(B) \leq \lambda_{\max}(q)\sigma_{\max}^2(a). \quad (\text{II.2})$$

Proof: For the upper bound, note that for any vector v ,

$$v^T a^T q a v \leq \lambda_{\max}(q) \|a v\|^2 \leq \lambda_{\max}(q) \sigma_{\max}^2(a) \|v\|^2.$$

Applying this inequality to a unit eigenvector for the maximum eigenvalue of c gives the result. The lower bound is obtained analogously:

$$v^T a^T q a v \geq \lambda_{\min}(q) \|a v\|^2 \geq \lambda_{\min}(q) \sigma_{\min}^2(a) \|v\|^2.$$

Applying this inequality to a unit eigenvector for the minimum eigenvalue of c completes the proof. ■

From this theorem, we get a simple bound on the condition number of $\kappa(B)$:

$$\kappa(B) = \frac{\lambda_{\max}(B)}{\lambda_{\min}(B)} \leq \frac{\lambda_{\max}(q)\sigma_{\max}^2(a)}{\lambda_{\min}(q)\sigma_{\min}^2(a)}. \quad (\text{II.3})$$

Since we typically have bounds on the eigenvalues of q , all that remains is to characterize the singular values of a in terms of the individual a_k . This is done in the next result which uses the relation

$$a^T a = \begin{pmatrix} I + a_2^T a_2 & a_2^T & 0 & \cdots \\ a_2 & I + a_3^T a_3 & & \vdots \\ \vdots & & \ddots & a_N^T \\ 0 & a_N & I + a_{N+1}^T a_{N+1} \end{pmatrix} \quad (\text{II.4})$$

where we define $a_{N+1} := 0$, so that the bottom right entry is the identity matrix.

Theorem 2.2: The following bounds hold for the singular values of a :

$$\begin{aligned} \min_k \{1 + \sigma_{\min}^2(a_{k+1}) - \sigma_{\max}(a_k) - \sigma_{\max}(a_{k+1})\} \\ \leq \sigma_{\min}^2(a^T a) \leq \sigma_{\max}^2(a^T a) \leq \\ \max_k \{1 + \sigma_{\max}^2(a_{k+1}) + \sigma_{\max}(a_k) + \sigma_{\max}(a_{k+1})\} \end{aligned} \quad (\text{II.5})$$

Proof: Let $v = \text{vec}(\{v_1, \dots, v_N\})$ be any eigenvector of $a^T a$, so that

$$a^T a v = \lambda v. \quad (\text{II.6})$$

Without loss of generality, suppose that the v_k component has largest norm out of $[1, \dots, N]$. Then from the k th block of (II.6), we get

$$a_k v_{k-1} + (I + a_{k+1}^T a_{k+1}) v_k + a_{k+1}^T v_{k+1} = \lambda v_k. \quad (\text{II.7})$$

Let $u_k = \frac{v_k}{\|v_k\|}$. Multiplying (II.7) on the left by v_k^T , dividing by $\|v_k\|^2$, and rearranging terms, we get

$$\begin{aligned} 1 + u_k^T a_{k+1}^T a_{k+1} u_k - \lambda = -u_k^T a_k \frac{v_{k-1}}{\|v_k\|} - u_k^T a_{k+1}^T \frac{v_{k+1}}{\|v_k\|} \\ \leq \sigma_{\max}(a_k) + \sigma_{\max}(a_{k+1}). \end{aligned} \quad (\text{II.8})$$

This relationships in (II.8) yield the upper bound

$$\lambda \leq 1 + \sigma_{\max}^2(a_{k+1}) + \sigma_{\max}(a_k) + \sigma_{\max}(a_{k+1}) \quad (\text{II.9})$$

and the lower bound

$$\begin{aligned} \lambda \geq 1 + u_k^T a_{k+1}^T a_{k+1} u_k - \sigma_{\max}(a_k) - \sigma_{\max}(a_{k+1}) \\ \geq 1 + \sigma_{\min}^2(a_{k+1}) - \sigma_{\max}(a_k) - \sigma_{\max}(a_{k+1}). \end{aligned} \quad (\text{II.10})$$

Taking the minimum over k in the lower bound and maximum over k for the upper bound completes the proof. ■

Corollary 2.1: Let v^{\min} be the eigenvector corresponding to $\lambda_{\min}(a^T a)$, and suppose that $\|v_k^{\min}\|$ is the component with the largest norm. Then we have the lower bound

$$\lambda_{\min}(a^T a) \geq 1 + \sigma_{\min}^2(a_{k+1}) - \sigma_{\max}(a_k) - \sigma_{\max}(a_{k+1}). \quad (\text{II.11})$$

In particular, since $a_{N+1} = 0$,

$$\lambda_{\min}(a^T a) \geq 1 - \sigma_{\max}(a_N). \quad (\text{II.12})$$

The bound II.12 reveals the vulnerability of the system $a^T a$ to the behavior of the last component.

For Kalman smoothing problems, the matrix $a^T q a$ corresponds to $G^T Q^{-1} G$. The components a_k correspond to the process models G_k . These components are often *identical* for all k , or they are all constructed from ODE discretizations, so that their singular values are similarly behaved across k . By (II.11), we see that the last component emerges as the weakest link, since regardless of how well the G_k are behaved for $k = 2, \dots, N-1$, the condition number can go to infinity if any singular value for G_N is larger than 1. The last block can thus have a very powerful effect on the entire system.

III. FORWARD BLOCK THOMAS ALGORITHM

We now present the forward block Thomas algorithm. *Forward* here refers to the fact that the algorithm starts and ends in the top left corner, in other words, it goes forward and then backward. In contrast, the *backward* block Thomas algorithm we present in Section V will start and end in the bottom right corner, so it goes backward, then forward.

Suppose for $k = 1, \dots, N$, $b_k \in \mathbf{R}^{n \times n}$, $e_k \in \mathbf{R}^{n \times \ell}$, $r_k \in \mathbf{B}^{n \times \ell}$, and for $k = 2, \dots, N$, $c_k \in \mathbf{R}^{n \times n}$. We define the corresponding block tridiagonal system of equations

$$\begin{pmatrix} b_1 & c_2^T & 0 & \cdots & 0 \\ c_2 & b_2 & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & \cdots & c_{N-1} & b_{N-1} & c_N^T \\ 0 & \cdots & 0 & c_N & b_N \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N-1} \\ e_N \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_{N-1} \\ r_N \end{pmatrix} \quad (\text{III.1})$$

For positive definite systems, the forward block Thomas algorithm is defined as follows [8, algorithm 4]:

Algorithm 3.1: The inputs to this algorithm are $\{c_k\}_{k=2}^N$, $\{b_k\}_{k=1}^N$, and $\{r_k\}_{k=1}^N$ where each $c_k \in \mathbf{R}^{n \times n}$, $b_k \in \mathbf{R}^{n \times n}$, and $r_k \in \mathbf{R}^{n \times \ell}$. The outputs are a sequence $\{e_k\}_{k=1}^N$ and η where each $e_k \in \mathbf{R}^{n \times \ell}$ and $\eta \in \mathbf{R}$. The sequence solves equation (III.1) and the λ that is equal to the log of the determinant of the block tridiagonal matrix in equation (III.1).

- 1) Set $d_1 = b_1$. Set $s_1 = r_1$.
- 2) For $k = 2$ To N :
 - Set $d_k = b_k - c_k d_{k-1}^{-1} c_k^T$.
 - Set $s_k = r_k - c_k d_{k-1}^{-1} s_{k-1}$.
- 3) Set $e_N = d_N^{-1} s_N$.
- 4) For $k = N-1$ To 1 :
 - Set $e_k = d_k^{-1} (s_k - c_{k+1}^T e_{k+1})$.
- 5) Set $\eta = \sum_{k=1}^N \log \det(d_k)$.

Before we discuss stability results for this algorithm (see theorem 3.2), we prove that the RTS smoother is an implementation of this algorithm for matrix C in (I.10).

Theorem 3.1: When applied to C in (I.10), Algorithm 3.1 is equivalent to the Rauch-Tung-Striebel [23] smoother.

Proof: Looking at the very first block, we substitute in the Kalman data structures (I.11) into step 2 of Algorithm 3.1. Writing this step requires introducing some structures which may be familiar to the reader from Kalman filtering literature.

$$\begin{aligned} P_{1|1}^{-1} &:= Q_1^{-1} + H_1^\top R_1^{-1} H_1 \\ P_{2|1}^{-1} &:= (G_1 P_{1|1} G_1^\top + Q_2)^{-1} \\ &= Q_2^{-1} - (Q_2^{-1} G_2)^\top (P_{1|1}^{-1} + G_2^\top Q_2^{-1} G_2)^{-1} (Q_2^{-1} G_2) \quad (\text{III.2}) \\ P_{2|2}^{-1} &:= P_{2|1}^{-1} + H_2^\top R_2^{-1} H_2 \\ d_2 &= c_2 - a_2^\top d_1^{-1} a_2 = P_{2|2}^{-1} + G_3^\top Q_3^{-1} G_3 \end{aligned}$$

These relationships can be seen quickly from [2, Theorem 2.2.7]. The matrices $P_{k|k}$, $P_{k|k-1}$ often appear in Kalman literature: they represent covariances of the state at time k given the the measurements $\{z_1, \dots, z_k\}$, and the covariance of the a priori state estimate at time k given measurements $\{z_1, \dots, z_{k-1}\}$, respectively.

The key fact from III.2 is that

$$d_2 = P_{2|2}^{-1} + G_3^\top Q_3^{-1} G_3.$$

Using the same computation for the generic tuple $(k, k+1)$ rather than $(1, 2)$ establishes

$$d_k = P_{k|k}^{-1} + G_{k+1}^\top Q_{k+1}^{-1} G_{k+1}. \quad (\text{III.3})$$

We now apply this technique to the right hand side $r = H^\top R^{-1} z + G^\top Q^{-1} w$. We have

$$\begin{aligned} y_{2|1} &:= (Q_2^{-1} G_2)^\top (P_{1|1}^{-1} + G_2^\top Q_2^{-1} G_2)^{-1} (H_1^\top R_1^{-1} z_1 + G_1^\top P_{0|0}^{-1} x_0) \\ y_{2|2} &:= H_2^\top R_2^{-1} z_2 + y_{2|1} \\ s_2 &= r_2 - a_2^\top d_1^{-1} r_1 = y_{2|2} \end{aligned} \quad (\text{III.4})$$

These relationships also follow from [2, Theorem 2.2.7]. The quantities $y_{2|1}$ and $y_{2|2}$ may be familiar to the reader from the information filtering literature: they are preconditioned estimates

$$y_{k|k} = P_{k|k}^{-1} x_{k|k}, \quad y_{k|k-1} = P_{k|k-1}^{-1} x_{k|k-1}, \quad (\text{III.5})$$

where $x_{k|k}$ is the estimate of x_k given $\{z_1, \dots, z_k\}$ and

$$x_{k|k-1} = G_k x_{k-1|k-1}$$

is the best prediction of the state x_k given $\{z_1, \dots, z_{k-1}\}$.

Again, by applying the computation to k , we have precisely that $s_k = y_{k|k}$. From these results, it immediately follows that e_N computed in step 3 of Algorithm 3.1 is the Kalman filter estimate (and the RTS smoother estimate) for time point N (see (III.3)):

$$e_N = d_N^{-1} s_N = (P_{N|N}^{-1} + 0)^{-1} P_{N|N}^{-1} x_{N|N} = x_{N|N}. \quad (\text{III.6})$$

We turn now to Step 4 of Algorithm 3.1. First, following [23, (3.29)], we define

$$C_k = P_{k|k} G_{k+1}^\top P_{k+1|k}^{-1}. \quad (\text{III.7})$$

To save space, we also use shorthand

$$\hat{P}_k := P_{k|k}, \quad \hat{x}_k := x_{k|k}. \quad (\text{III.8})$$

At the first step, we obtain

$$\begin{aligned} e_{N-1} &= d_{N-1}^{-1} (s_{N-1} - c_N^\top e_N) \\ &= (\hat{P}_{N-1}^{-1} + G_N^\top Q_N^{-1} G_N)^{-1} (\hat{P}_{N-1}^{-1} \hat{x}_{N-1} - G_N^\top Q_N^{-1} \hat{x}_N) \\ &= (\hat{P}_{N-1}^{-1} + G_N^\top Q_N^{-1} G_N)^{-1} \hat{P}_{N-1}^{-1} \hat{x}_{N-1} - C_{N-1} \hat{x}_N \\ &= \hat{x}_{N-1} - C_{N-1} (G_N x_{N-1} - \hat{x}_N) \\ &= x_{N-1|N-1} + C_{N-1} (x_{N|N} - G_N x_{N-1|N-1}), \end{aligned} \quad (\text{III.9})$$

where the Woodbury inversion formula was used to get from line 3 to line 4. Comparing this to [23, (3.28)], we find that $e_{N-1} = x_{N-1|N}$, i.e. the Rauch-Tung-Striebel smoothed estimate. The computations above, when applied to the general tuple $(k, k+1)$ instead of $(N-1, N)$, show that every e_k is equivalent to $x_{k|N}$, which completes the proof. ■

In 1965, Rauch, Tung and Striebel showed that their smoother solves the maximum likelihood problem for $\mathbf{p}(\{x_k\}|\{z_k\})$ [23], which is equivalent to (I.5). Theorem 3.1 adds to this understanding, showing that it is *precisely* the forward block Thomas algorithm. Moreover, the proof explicitly shows how the Kalman filter estimates $x_{k|k}$ can be obtained as the forward block Thomas proceeds to solve system (I.9). For efficiency, we did not include any expressions related to the Kalman gain; the interested reader can find these relationships in [2, Chapter 2]. They emerge through application of the Woodbury matrix identity to the equations we presented.

We now turn our attention to the stability of the forward Thomas algorithm for block tridiagonal systems. One such analysis appears in [8, Lemma 6], and has been used frequently to justify the forward Thomas as the method of choice in many Kalman smoothing applications. Below, we review this result, and show that it has a critical flaw precisely for Kalman smoothing systems.

Lemma 3.1: Suppose we are given sequences $\{q_k\}_{k=0}^N$, $\{a_k\}_{k=2}^N$, and $\{u_k\}_{k=1}^N$, where each $q_k \in \mathbf{R}^{n \times n}$ is symmetric positive definite, each $u_k \in \mathbf{R}^{n \times n}$ is symmetric semi-positive definite, each $a_k \in \mathbf{R}^{n \times n}$. Define $b_k \in \mathbf{R}^{n \times n}$ by

$$b_k = u_k + q_k^{-1} + a_{k+1}^\top q_{k+1}^{-1} a_{k+1}, \quad \text{where } k = 1, \dots, N$$

Define $c_k \in \mathbf{R}^{n \times n}$ by

$$c_k = q_k^{-1} a_k, \quad \text{where } k = 2, \dots, N$$

Suppose there is an $\alpha > 0$ such that all the eigenvalues of $a_k^\top q_k^{-1} a_k$ are greater than or equal α . Suppose there is a $\beta > 0$ such that all the eigenvalues of b_k are less than or equal β . Further suppose we execute Algorithm 3.1 with corresponding input sequences $\{b_k\}_{k=1}^N$ and $\{c_k\}_{k=2}^N$. It follows that each d_k generated by the algorithm is symmetric positive definite and has condition number less than or equal β/α .

To understand what can go wrong with this algorithm, consider b_N in the Kalman smoothing context where the corresponding matrix entries are given by [8, equations (12) and (13)]. The matrix b_N in Lemma 3.1 is given by

$$b_N = u_N + q_N^{-1} + a_{N+1}^T q_{N+1}^{-1} a_{N+1}$$

where the correspondence to [8, (13)] is given by b_N corresponds to B_N , q_N corresponds to Q_N , a_N corresponds to G_N , and u_N corresponds to

$$H_N^T R_N^{-1} H_N.$$

In the context of [8], G_{k+1} is the model for the next state vector and at $k = N$ there is no next state vector. Hence, $G_{N+1} = 0$. Thus, in the context of Lemma 3.1, $a_N = 0$ and hence $\alpha = 0$ which contradicts the Lemma assumptions.

Does this mean that the forward block Thomas algorithm (and hence the Kalman filter and RTS smoother) are inherently numerically unstable, unless they have measurements to stabilize them? It turns out that this is not the case; in fact we can prove a powerful theorem that relates numerical stability of the forward block Thomas algorithm to the condition number of the system (II.1), already characterized in Theorem 2.1.

Theorem 3.2: Consider any block tridiagonal system $A \in \mathbb{R}^N$ of form (III.1). and suppose we are given a lower bound α_L and an upper bound α_U on the eigenvalues of this system:

$$0 < \alpha_L \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \alpha_U. \quad (\text{III.10})$$

If we apply the block Thomas iteration

$$d_k = b_k - c_k d_{k-1}^{-1} c_k^T,$$

then

$$0 < \alpha_L \leq \lambda_{\min}(d_k) \leq \lambda_{\max}(d_k) \leq \alpha_U \quad \forall k. \quad (\text{III.11})$$

In other words, the block Thomas iteration preserves eigenvalue bounds (and hence the condition number) for each block, and hence will be numerically stable when the full system is well conditioned.

Proof: For simplicity, we will focus only on the lower bound, since the same arguments apply for the upper bound. Note that $b_1 = d_1$, and the eigenvalues of d_1 must satisfy

$$\alpha_L \leq \lambda_{\min}(d_1)$$

since otherwise we can produce an unit-norm eigenvector $v_1 \in \mathbb{R}^n$ of d_1 with $v_1^T d_1 v_1 < \alpha_L$, and then form the augmented unit vector $\tilde{v}_1 \in \mathbb{R}^N$ with v_1 in the first block, and every other entry 0. Then we have

$$\tilde{v}_1^T A \tilde{v}_1 < \alpha_L,$$

which violates (V.2). Next, note that

$$S_1 A S_1^T = \begin{pmatrix} b_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & & c_{N-1} & b_{N-1} & c_N^T \\ 0 & \cdots & 0 & c_N & b_N \end{pmatrix} \quad (\text{III.12})$$

where

$$d_2 = b_2 - c_2 d_1^{-1} c_2^T$$

and

$$S_1 = \begin{pmatrix} I & 0 & 0 & \cdots & 0 \\ -c_2 d_1^{-1} & I & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & & 0 & I & 0 \\ 0 & \cdots & 0 & 0 & I \end{pmatrix}.$$

Suppose now that d_2 has an eigenvalue that is less than α_L . Then we can produce a unit eigenvector v_2 of d_2 with $v_2^T d_2 v_2 < \alpha_L$, and create an augmented unit vector

$$\tilde{v}_2 = [0_{1 \times n} \quad v_2^T \quad 0_{1 \times n(N-2)}]^T$$

which satisfies

$$\tilde{v}_2^T S_1 A S_1^T \tilde{v}_2 < \alpha_L. \quad (\text{III.13})$$

Next, note that

$$\hat{v}_2^T := \tilde{v}_2^T S_1 = [-v_2^T c_2 d_1^{-1} \quad v_2^T \quad 0_{1 \times n(N-2)}]^T,$$

so in particular $\|\hat{v}_2\| \geq 1$. From (III.13), we now have

$$\hat{v}_2^T A \hat{v}_2 \leq \alpha_L \leq \alpha_L \|\hat{v}_2\|,$$

which violates (V.2). To complete the proof, note that the lower $n(N-1) \times n(N-1)$ block of $S_1 A S_1^T$ has the same form as A , with (V.2) holding for this modified system. The reduction technique can now be repeatedly applied. ■

Note that Theorem 3.2 applies to *any* block tridiagonal system. When applied to the Kalman smoothing setting, if the system $G^T Q^{-1} G$ is well-conditioned, we know that the forward block Thomas (and hence the Kalman filter and RTS smoother) will behave well for any measurement model. Moreover, even if $G^T Q^{-1} G$ has a bad condition number, it is possible that the measurement term $H^T R^{-1} H$ (see (I.9)) will improve the condition number. More general Kalman smoothing applications may not have this advantage. For example, the initialization procedure in [5] requires the inversion of systems analogous to $G^T Q^{-1} G$.

IV. INVERTIBLE MEASUREMENT COMPONENT

We now return to the system (I.9), and consider the case where $H^T R^{-1} H$ is an invertible matrix. Note that this is not true in general, and in fact our analysis, as applied to the Kalman smoothing problem, did not use any assumptions on this term. However, if we know that

$$F := H^T R^{-1} H \quad (\text{IV.1})$$

is an invertible matrix, then we can consider an alternative approach to solving (I.9). Applying the Woodbury inversion formula, we obtain

$$(G^T Q^{-1} G + F)^{-1} = F^{-1} - F^{-1} G^T (Q + G F^{-1} G^T)^{-1} G F^{-1} \quad (\text{IV.2})$$

Now, the solution to (I.9) can be found by applying this explicit inverse to the right hand side

$$\text{rhs} := H^T R^{-1} z + G^T Q^{-1} \zeta$$

and the key computation becomes

$$(Q + G F^{-1} G^T) x = G F^{-1} \text{rhs}. \quad (\text{IV.3})$$

Note that the matrix $Q + GF^{-1}G^T$ is block tridiagonal, since Q, F^{-1} are block diagonal, and G is lower block bidiagonal. Therefore, we have reduced the problem to a system of the form (III.1). Moreover, at a glance we can see the lower eigenvalues of this system are bounded below by the eigenvalues of Q , while upper bounds can be constructed from eigenvalues of Q, G_k and F^{-1} . Under very mild conditions, this system can be solved in a stable way by the forward tridiagonal algorithm, which would also give a modified filter and smoother. This is not surprising, since we have assumed the extra hypothesis that F is invertible.

V. BACKWARD BLOCK TRIDIAGONAL ALGORITHM

Now that we have abstracted Kalman smoothing problems and algorithms to solutions of block tridiagonal systems, we can consider novel techniques for their solution. In this section, we propose a *backward* block Thomas algorithm, and then show it has a very interesting property when applied to the Kalman smoothing setting. Recall that *backward* means the algorithm starts and ends in the lower right hand corner, so it first goes backward, and then forward.

To derive the backward algorithm, we begin with system (III.1). Subtracting $c_N^T b_N^{-1}$ times row N from row $N-1$, we obtain the following equivalent equation:

$$\begin{pmatrix} b_1 & c_2^T & 0 & \cdots & 0 \\ c_2 & b_2 & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & b_{N-2} & & c_{N-1}^T & 0 \\ 0 & c_{N-1} & b_{N-1} - c_N^T b_N^{-1} c_N & & 0 \\ 0 & \cdots & 0 & c_N & b_N \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N-1} \\ e_N \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_{N-1} - c_N^T b_N^{-1} r_N \\ r_N \end{pmatrix}.$$

We iterate this procedure until we reach the first row of the matrix, using d_k to denote the resulting diagonal blocks, and s_k the corresponding right hand side of the equations; i.e.,

$$d_N = b_N, \quad d_k = b_k - c_{k+1}^T d_{k+1}^{-1} c_{k+1} \quad (k = N-1, \dots, 1) \\ s_N = e_N, \quad s_k = r_k - c_{k+1}^T d_{k+1}^{-1} s_{k+1} \quad (k = N-1, \dots, 1).$$

We obtain the following lower triangular system:

$$\begin{pmatrix} d_1 & 0 & \cdots & 0 \\ c_2 & d_2 & 0 & \cdots & 0 \\ & & \ddots & & \vdots \\ \vdots & & & c_{N-1} & d_{N-1} & 0 \\ 0 & & & c_N & b_N \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{N-1} \\ e_N \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N-1} \\ r_N \end{pmatrix} \quad (\text{V.1})$$

Now we can solve for the *first* block vector and then proceed back down, doing back substitution. We thus obtain the following algorithm:

Algorithm 5.1: A block tridiagonal extension of the backward Thomas algorithm: The inputs to this algorithm are $\{c_k\}$, $\{b_k\}$, and $\{r_k\}$. The output is a sequence $\{e_k\}$ that solves equation (III.1) the η which is equal to the log of the determinant of the block tridiagonal matrix in equation (III.1).

- 1) Set $d_N = b_N$ and $s_N = r_N$.
- 2) For $k = N-1, \dots, 1$, set $d_k = b_k - c_{k+1}^T d_{k+1}^{-1} c_{k+1}$, $s_k = r_k - c_{k+1}^T d_{k+1}^{-1} s_{k+1}$.
- 3) Set $e_1 = d_1^{-1} s_1$.
- 4) For $k = 2, \dots, N$, set $e_k = d_k^{-1} (s_k - c_k e_{k-1})$.
- 5) Set $\eta = \sum_{k=1}^N \log \det(d_k)$.

First, we show that the backward block Thomas algorithm has the same numerical stability result as the forward block Thomas algorithm.

Theorem 5.1: Consider any block tridiagonal system $A \in \mathbb{R}^N$ of form (III.1). and suppose we are given the bounds α_L and α_U for the lower and upper bounds of the eigenvalues of this system

$$0 < \alpha_L \leq \lambda_{\min}(A) \leq \lambda_{\max}(A) \leq \alpha_U. \quad (\text{V.2})$$

If we apply the backward block Thomas iteration

$$d_k = b_k - c_{k+1}^T d_{k+1}^{-1} c_{k+1}$$

then we have

$$0 < \alpha_L \leq \lambda_{\min}(d_k) \leq \lambda_{\max}(d_k) \leq \alpha_U \quad \forall k. \quad (\text{V.3})$$

Proof: Note first that $d_N = b_N$, and satisfies (V.3) by the same argument as in the proof of Theorem 5.1. Define

$$S_N = \begin{pmatrix} I & 0 & 0 & \cdots & 0 \\ 0 & I & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & 0 & & I & 0 \\ 0 & \cdots & 0 & -c_N d_{N-1}^{-1} & I \end{pmatrix}.$$

and note that

$$S_N^T A S_N = \begin{pmatrix} d_1 & c_2^T & \cdots & 0 \\ c_2 & d_2 & c_3^T & \cdots & 0 \\ 0 & & \ddots & c_{N-1}^T & \vdots \\ \vdots & & c_{N-1} & d_{N-1} & 0 \\ 0 & \cdots & 0 & 0 & d_N \end{pmatrix}$$

Now an analogous proof to that of Theorem 5.1 can be applied to show the upper $n(N-1) \times n(N-1)$ block of $S_N^T A S_N$ satisfies (V.3). Applying this reduction iteratively completes the proof. ■

Theorems 3.2 and 5.1 show that both forward and backward Thomas algorithms are numerically stable when the block tridiagonal systems they are applied to are well conditioned. However, a different analysis can also be done for a particular class of block tridiagonal systems. This result, which applies to Kalman smoothing systems, shows that *the backward algorithm can behave stably even when the tridiagonal system has null singular values.*

For $v \in \mathbb{R}^{n \times n}$ we use the notation $|v|$ for the operator norm of the matrix v ; i.e.,

$$|v| = \sup\{|vw| : w \in \mathbb{R}^n, |w| = 1\}$$

Theorem 5.2: Suppose that the matrices c_k and b_k are given by

$$c_k = -q_k^{-1} a_k \\ b_k = q_k^{-1} + u_k + a_{k+1}^T q_{k+1}^{-1} a_{k+1}$$

where each $q_k \in \mathbf{R}^{n \times n}$ is positive definite, each $u_k \in \mathbf{R}^{n \times n}$ is positive semi-definite and $a_{N+1} = 0$. It follows that $d_k - q_k^{-1}$ is positive semi-definite for all k . Furthermore, if α is a bound for $|q_k|$, $|q_k^{-1}|$, $|u_k|$, and $|a_k|$, Then the condition number of d_k is bounded by $\alpha^2 + \alpha^6$.

Proof: We note that $d_N = q_N^{-1} + u_N$ so this conditions bound holds for $k = N$. Furthermore $d_N - q_N^{-1} = u_N$, so positive semi-definite assertion holds for $k = N$.

We now complete the proof by induction; i.e., suppose $d_{k+1} - q_{k+1}^{-1}$ is positive semi-definite

$$\begin{aligned} d_k &= b_k - c_{k+1}^T d_{k+1}^{-1} c_{k+1} \\ d_k - q_k^{-1} &= u_k + a_{k+1}^T q_{k+1}^{-1} a_{k+1} - a_{k+1}^T q_{k+1}^{-1} d_{k+1}^{-1} q_{k+1}^{-1} a_{k+1} \\ &= u_k + a_{k+1}^T q_{k+1}^{-1} [q_{k+1} - d_{k+1}^{-1}] q_{k+1}^{-1} a_{k+1} \end{aligned}$$

The assumption that $d_{k+1} - q_{k+1}^{-1}$ is positive semi-definite implies that that $q_{k+1} - d_{k+1}^{-1}$ is positive semi-definite. It now follows that $d_k - q_k^{-1}$ is the sum of positive semi-definite matrices and hence is positive semi-definite which completes the induction and hence proves that $d_k - q_k^{-1}$ is positive semi-definite.

We now complete the proof by should the condition number bound holds for index k . Using the last equation above, we have

$$\begin{aligned} |d_k| &\leq |u_k| + |a_{k+1}^T|^2 |q_{k+1}^{-1}|^2 |q_{k+1}| \\ &\leq \alpha + \alpha^5 \end{aligned}$$

Hence the maximum eigenvalue of d_k is less than or equal $\alpha + \alpha^5$. In addition, since $d_k - q_k^{-1}$ is positive semi-definite, the minimum eigen-value of d_k is greater than or equal the minimum eigen-value of q_k^{-1} , which is equal to the reciprocal of the maximum eigen-value of q_k . Thus the minimum eigen value of d_k is greater than or equal $1/\alpha$. Thus the condition number of d_k is bounded by $\alpha^2 + \alpha^6$ which completes the proof. ■

VI. NUMERICAL EXAMPLE

To put all of these ideas in perspective, we consider a toy numerical example. Let $n = 1$ and $N = 3$, let $q_k = 1$ for all k , and let $a_k = 120$. Then the system $a^T q a$ in (II.1) is given by

$$\begin{bmatrix} 14401 & 120 & 0 \\ 120 & 14401 & 120 \\ 0 & 120 & 1 \end{bmatrix},$$

and its minimum eigenvalue is 4.8×10^{-9} . To understand what goes wrong, we first note that the minimum and maximum eigenvalues of all the a_k 's except the last one coincide in this case, so the general condition of Corollary 2.1 will hold everywhere except at the last coordinate.

Now that we suspect the last coordinate, we can check the eigenvector corresponding to the minimum eigenvalue:

$$v^{\min} = [0.001 \quad -.008 \quad 1]^T.$$

Indeed, the component of the eigenvector with the largest norm occurs precisely in the last component, so we are in the case described by Corollary 2.1. In order to control the condition number of this system, one option is to make the (3,2) and

(2,3) coordinates less than 1 in absolute value. This *changes the original system*, but only at the last time point. Suppose we decide this is acceptable, and take these values to be 0.9 instead of 120. The new system has lowest eigenvalue 1. If we expand the system (by increasing the size of $\{a_k\}$, $\{q_k\}$, we find this stabilization technique works regardless of the size, since the last component is always the weakest link when all a_k are equal.

Next, suppose we apply the forward Thomas algorithm to the original (badly conditioned) system. We would get blocks

$$d_1 = 14401, \quad d_2 = 14400, \quad d_3 = 4.8222 \times 10^{-9}.$$

This toy example shows Theorem 3.2 in action — indeed, the eigenvalues of the blocks are bounded above and below by the eigenvalues of the full system. Unfortunately, in this case this leads to a terrible condition number, since we are working with an ill-conditioned system to begin with.

Now, suppose we apply the backward Thomas algorithm. We will get the following blocks:

$$d_3 = 1, \quad d_2 = 1, \quad d_1 = 1.$$

Note that even though the *blocks* are stable, this does not mean that the backward algorithm can accurately solve the system $a^T q a x = b$. In practice, it may perform better or worse than the forward algorithm for ill-conditioned systems.

Nonetheless, these results show the backward block Thomas has an important advantage: if it must be applied many times in an iterative context, it may happen that systems are badly conditioned at initial iterations, but the conditioning improves later on. In this context, the backward block Thomas it will run where the forward block Thomas will not. This was exactly the case in [5], where the introduction of the backward block Thomas stabilized the initialization procedure; for later iterations, the measurement terms came in for later iterations to improve the conditioning number of the system.

VII. CONCLUSIONS

In this paper, we have characterized the condition number of tridiagonal systems that arise in Kalman smoothing, see theorems 2.1 and 2.2. The analysis revealed that last block of the system can have a very strong effect on the system overall. In the Kalman smoothing context, it is important to note that the numerical stability conditions in theorems 2.1 and 2.2 do not require a_k in (II.1) to be invertible. In fact, they may be singular, which means the process matrices G_k , or derivatives of nonlinear process functions $g_k^{(1)}$, do not have to be invertible.

We then showed that any well-conditioned symmetric block tridiagonal system can be solved in a numerically stable manner by the forward block Thomas algorithm. Applied to the Kalman smoothing context, we explicitly showed that the forward block Thomas algorithm is equivalent to the RTS smoother. Finally, we introduced a new method, the backward block Thomas algorithm, which shares the numerical stability properties of the forward block Thomas algorithm, and can remain stable even when applied to ill-conditioned systems.

This property makes the backward Thomas algorithm preferable in an iterative context, where on occasion one may run into badly conditioned systems. The computational effort of the backward algorithm is equivalent to that of the forward; both scale linearly with the size of the tridiagonal system.

Taken together, these results provide insight into both numerical stability and algorithms for Kalman smoothing systems. Block tridiagonal systems arise both in the classic setting, and also in many new methods, that can handle nonlinear process/measurement models, robust penalties, and inequality constraints.

REFERENCES

- [1] B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, N.J., USA, 1979.
- [2] A. Aravkin. *Robust Methods with Applications to Kalman Smoothing and Bundle Adjustment*. PhD thesis, University of Washington, Seattle, WA, June 2010.
- [3] A. Aravkin, J. Burke, and G. Pillonetto. Robust and trend following kalman smoothers using student's t. In *Proc. of SYSID 2012*, 2012.
- [4] Aleksandr Y Aravkin, James V. Burke, and Gianluigi Pillonetto. Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory, 2013. *Submitted to Journal of Machine Learning Theory (JMLR)*.
- [5] A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1 -Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 56(12), 2011.
- [6] Yaakov Bar-Shalom, X. Rong Li, and Thiagalingam Kirubarajan. *Estimation with Applications to Tracking and Navigation*. John Wiley and Sons, 2001.
- [7] B. M. Bell, J. V. Burke, and G. Pillonetto. An inequality constrained nonlinear Kalman-Bucy smoother by interior point likelihood maximization. *Automatica*, 45(1):25–33, 2009.
- [8] B.M. Bell. The marginal likelihood for parameters in a discrete Gauss-Markov process. *IEEE Transactions on Signal Processing*, 48(3):870–873, 2000.
- [9] Charles Chui and Guanrong Chen. *Kalman Filtering*. Springer, 2009.
- [10] T. Cipra and R. Romera. Kalman filter with outliers and missing observations. *Sociedad de Estadística e Investigación Operativa*, 6(2):379–395, 1997.
- [11] Z.M. Durovic and B.D. Kovachevic. Robust estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 44(6):1292–1296, June 1999.
- [12] L. Fahrmeir and H. Kaufmann. On Kalman filtering, posterior mode estimation, and Fisher scoring in dynamic exponential family regression. *Metrika*, pages 37–60, 1991.
- [13] S. Farahmand, G.B. Giannakis, and D. Angelosante. Doubly robust smoothing of dynamical processes via outlier sparsity constraints. *IEEE Transactions on Signal Processing*, 59:4529–4543, 2011.
- [14] Mohinder S. Grewal, Lawrence R. Weill, and Angus P. Andrews. *Global Positioning Systems, Inertial Navigation, and Integration*. John Wiley and Sons, 2nd edition, 2007.
- [15] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the AMSE - Journal of Basic Engineering*, 82(D):35–45, 1960.
- [16] R. E. Kalman and R. S. Bucy. New results in linear filtering and prediction theory. *Trans. ASME J. Basic Eng.*, 83:95–108, 1961.
- [17] S.A. Kassam and H.V. Poor. Robust techniques for signal processing: A survey. *Proceedings of the IEEE*, 73(3):433–481, March 1985.
- [18] R.L. Kirlin and A. Moghaddamjoo. Robust adaptive Kalman filtering for systems with unknown step inputs and non-Gaussian measurement errors. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34(2):252–263, March 1986.
- [19] G. Kitagawa and Will Gersch. A smoothness priors time-varying ar coefficient modeling of nonstationary covariance time series. *Automatic Control, IEEE Transactions on*, 30(1):48–56, Jan.
- [20] C.J. Masreliez and R.D. Martin. Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE Transactions on Automatic Control*, AC-22(3):361–371, June 1977.
- [21] R.J. Meinhold and N.D. Singpurwalla. Robustification of Kalman filter models. *Journal of the American Statistical Association*, 84(406):479–486, June 1989.
- [22] Ian R. Petersen and Andrey V. Savkin. *Robust Kalman Filtering for Signals and Systems with Large Uncertainties*. Control Engineering. Birkhauser, 1999.
- [23] H. E. Rauch, F. Tung, and C. T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA J.*, 3(8):1145–1150, 1965.
- [24] I.C. Schick and S.K. Mitter. Robust recursive estimation in the presence of heavy-tailed observation noise. *The Annals of Statistics*, 22(2):1045–1080, June 1994.
- [25] S.J. Wright. Solution of discrete-time optimal control problems on parallel computers. *Parallel Computing*, 16:221–238, 1990.
- [26] S.J. Wright. Interior point methods for optimal control of discrete-time systems. *Journal of Optimization Theory and Applications*, 77:161–187, 1993.